Machine Learning Models for Classification of Lung Cancer and Selection of Genomic Markers Using Array Gene Expression Data

C.F. Aliferis[1], I. Tsamardinos[1], P.P. Massion[2], A. Statnikov[1], N. Fananapazir[1], D. Hardin[3]

[1]Department of Biomedical Informatics, [2]Department of Medicine, [3]Department of Mathematics,
Vanderbilt University, 2209 Garland Ave., Nashville, TN 37232-8340, USA
{constantin.aliferis, ioannis.tsamardinos, pierre.massion, alexander.statnikov, nafeh.fananapazir, douglas.p.hardin}@vanderbilt.edu

Abstract

This research explores machine learning methods for the development of computer models that use gene expression data to distinguish between tumor and non-tumor, between metastatic and non-metastatic, and between histological subtypes of lung cancer. A second goal is to identify small sets of gene predictors and study their properties in terms of stability, size, and relation to lung cancer. We apply four classifier and two gene selection algorithms to a 12,600 oligonucleotide array dataset from 203 patients and normal human subjects. The resulting models exhibit excellent classification performance. Gene selection methods reduce drastically the genes necessary for classification. Selected genes are very different among gene selection methods, however. A statistical method for characterizing the causal relevance of selected genes is introduced and applied.

Introduction and Problem Statement

Lung cancer is the third most common cancer in the United States yet causes more deaths than breast, colon and prostate cancer combined (Parker et al. 1996). In spite of recent advances in treatment, approximately 90% of the estimated 170,000 patients diagnosed with lung cancer in 2002 are expected to eventually die of their disease. Major goals of lung cancer research is to understand the molecular basis of disease, to offer patients with better early diagnostic and therapeutic tools, and to individualize therapeutics based on molecular determinants of the tumors. The present research addresses three aims related to creating clinically and biologically useful molecular models of lung cancer using gene expression data: (a) Apply supervised classification methods to construct computational models that distinguish between: (i) Cancerous vs Normal Cells; (ii) Metastatic vs Non-Metastatic cells; and (iii) Adenocarcinomas vs Squamous carcinomas. (b) Apply feature selection methods to reduce the number of gene markers such that small sets of genes can distinguish among the different states (and ideally reveal important genes in the pathophysiology of lung cancer). (c) Compare the performance of the machine learning (classifier and feature selection) methods employed in these dataset and tasks.

Data and Methods

Data. We analyzed the data of Bhattacharjee et al., which is a set of 12,600 gene expression measurements (Affymetrix oligonucleotide arrays) per patient from 203 patients and normal subjects. The original study explored identification of new molecular subtypes and their association to survival. Hence the experiments presented here do not replicate or overlap with those of (Bhattacharjee et al. 2001).

Classifiers. In our experiments we used linear and polynomial-kernel Support Vector Machines (LSVM, and PSVM respectively) (Scholkopf et al. 1999), K-Nearest Neighbors (KNN) (Duda et al. 2001), and feed-forward Neural Networks (NNs) (Hagan et al. 1996). For SVMs we used the LibSVM base implementation (Chang et al.), with C chosen from the set: {1e-14, 1e-3, 0.1, 1, 10, 100, 1000} and degree from the set {2, 3, 4}. For KNN, we chose k from the range [1,…,number_of_variables] using our own implementation of the algorithm. For NNs we used the Matlab Neural Network Toolbox (Demuth et al. 2001) with 1 hidden layer, number of units chosen (heuristically) from the set {2, 3, 5, 8, 10, 30, 50}, variable learning rate back propagation, performance goal=1e-8 (i.e., an arbitrary value very close to zero), a fixed momentum of 0.001, and number of epochs chosen from the range [100,…,10000]. The number of epochs in particular was optimised via special scripts with nested cross-validation during training such that training would stop when the error in an independent validation set would start increasing. To avoid overfitting, either in the sense of optimising parameters for classifiers, or in the sense of estimating final performance of the best classifier/gene set found (Duda et al. 2001) a nested cross-validation design was employed. In this design, the outer layer of cross-validation estimates the performance of the optimised classifiers while the inner layer chooses the best parameter configuration for each classifier). For the two tasks (adenocarcinoma-squamous, and normal-cancer) we used 5-fold cross-validation while for the metastatic-nonmetastatic task we used 7-fold cross-validation (since we had only 7 metastatic cases in the sample). To ensure optimal use of the available sample, we required that data splits were balanced (i.e., instances with the rarer of the two categories of each target would appear in the same proportion in each random data split).

Cancer vs normal				Adenocarcinomas vs squamous carcinomas			Metastatic vs non-metastatic adenocarcinomas		
classifiers	RFE	UAF	All Features	RFE	UAF	All Features	RFE	UAF	All Features
LSVM	97.03%	99.26%	99.64%	98.57%	99.32%	98.98%	96.43%	95.63%	96.83%
PSVM	97.48%	99.26%	99.64%	98.57%	98.70%	99.07%	**97.62%**	**96.43%**	96.33%
KNN	87.83%	97.33%	98.11%	91.49%	95.57%	97.59%	92.46%	89.29%	92.56%
NN	97.57%	**99.80%**	N/A	98.70%	**99.63%**	N/A	96.83%	86.90%	N/A
Averages over classifier	94.97%	**98.91%**	99.13%	96.83%	**98.30%**	98.55%	**95.84%**	92.06%	95.24%

Table 1. Classification Performance Of All Classifier/Gene Selection Method Employed

Feature Selection. The feature (or variable) selection problem can be stated as follows: *given a set of predictors ("features") V and a target variable T, find a minimum subset F of V that achieves maximum classification performance of T (relative to a dataset, task, and a set of classifier-inducing algorithms).* Feature selection is pursued for a number of reasons: for many practical classifiers it may improve performance; a classification algorithm may not scale up to the size of the full feature set either in sample or time; feature selection may allow researchers to better understand the domain; it may be cheaper to collect a reduced set of predictors; and, finally, it may be safer to collect a reduced set of predictors (Tsamardinos and Aliferis 2003). Feature selection methods are typically of the wrapper or the filter variety. Wrapper algorithms perform a heuristic search in the space of all possible feature subsets and evaluate each visited state by applying the classifier for which they intend to optimise the feature subset. Common examples of heuristic search are hill climbing (forward, backward, and forward-backward), simulated annealing, and Genetic Algorithms. The second class of feature selection algorithms is filtering. Filter approaches select features on the basis of statistical properties of their joint distribution with the target variable. We used two such methods:

(a) Recursive Feature Elimination (RFE). RFE builds on SVM classification. The basic procedure can be summarized as follows (Guyon et al. 2002):

1. Build a linear Support Vector Machine classifier using all V features
2. Compute weights of all features and choose the first |V|/2 features (sorted by weight in decreasing order)
3. Repeat steps #1 and #2 until one feature is left
4.Choose the feature subset that gives the best performance
5. Optional: Give the best feature set to other classifiers of choice.

RFE was employed using the parameters employed in (Guyon et al. 2002).

(b) Univariate Association Filtering (UAF). UAF examines the association of each individual predictor feature (gene) to the target variable. The procedure is common in applied classical statistics (Tabachnick et al. 1989) and can be summarized as follows:

1. Order all predictors according to strength of pair-wise (i.e., univariate) association with target
2. Choose the first k predictors and feed them to the classifier

We note that various measures of association may be used. In our experiments we use Fisher Criterion Scoring, since previous research has shown that this is an appropriate measure for gene expression data (Furey et al. 2000). In practice k is often chosen arbitrarily based on the limitations of some classifier relative to the available distribution and sample, or can be optimised via cross-validation (our chosen approach). We used our own implementations of RFE and UAF.

Performance Evaluation. In all reported experiments we used the area under the Receiver Operator Characteristic (ROC) curve (AUC) to evaluate the quality of the produced models (Provost, Fawcett and Kohavi 1998). Unlike accuracy (i.e., proportion of correct classifications) this metric is independent of the distribution of classes. It is also independent of the misclassification cost function. Since in the lung cancer domain such cost functions are not generally agreed upon, we chose to use the AUC metric. We note that by emphasizing robustness AUC also captures more the intrinsic quality of what has been learned (or is learnable) in the domain and in that sense can be considered more useful for biomedical discovery. We use our own Matlab implementation of computation of AUC using the trapezoidal rule (DeLong et al. 1998). Statistical comparisons among AUCs were performed using a paired Wilcoxon rank sum test (Pagano et al. 2000).

Results

Classification Performance. Table 1 shows the average cross-validated AUC performance of models built using all genes as well as genes selected by the RFE and UAF

Feature Selection Method	Number of features discovered		
	Cancer vs normal	Adenocarcinomas vs squamous carcinomas	Metastatic vs non-metastatic adenocarcinomas
RFE	6	12	**6**
UAF	**100**	**500**	500

Table 2. Parsimony of Gene Marker Sets (In Bold: Better-Performing Models)

	Cancer vs normal		Adenocarcinomas vs squamous carcinomas		Metastatic vs non-metastatic adenocarcinomas	
Contributed by method on the left compared with method on the right	RFE	UAF	RFE	UAF	RFE	UAF
RFE	0	2	0	5	0	2
UAF	96	0	493	0	496	0

Table 3. Relative Overlap of Genes Between the Gene Selection Methods For Each Task

	Cancer vs normal		Adenocarcinomas vs squamous carcinomas		Metastatic vs non-metastatic adenocarcinomas	
Percentage of genes from method on the left eliminated by genes from methods on the right	RFE	UAF	RFE	UAF	RFE	UAF
RFE		100.00%		100.00%		100.00%
UAF	0.00%		31.24%		7.46%	
p-value	**0.001**		**<0.001**		**<0.001**	

Table 4. Relative Conditional Blocking

methods. We see for example that UAF in combination with NNs gives the best-performing model (with almost perfect classification performance) in two out of three tasks and that KNN exhibits poorer performance compared to the other classifiers. The UAF method is the most robust across all classifier-inducing algorithms tested for two of three classification tasks. Boldface denotes best values such that non-bold fonts are statistically significantly different from the best values for each task at the 5% significance level. Note that NNs could not be run in reasonable time using the full set of genes.

Gene Selection Analysis. In Table 2 we show the numbers of genes selected by each method for each classification task. We note that these genes were selected by running the methods on all available sample after (cross-validated) analyses of Table 1 were completed. This contrasts parsimony with classification performance. RFE produces the most parsimonious gene marker sets. In boldface we denote the set that gives the best classifier for each task. Table 3 examines the relative overlap of selected genes among the gene selection methods for each classification task. As can be seen, the two methods contribute several additional genes with respect to each other (i.e., neither is redundant). Clearly the two gene selection methods exhibit a different inductive bias (i.e., a preference criterion for selected genes). *Characterizing this inductive bias is very important since it effectively answers the question "what is the biological meaning that this gene was selected by method X?"* Since we are fundamentally interested in causal discovery of gene-gene interactions in this domain, and as a step toward such an understanding, we apply the following analysis: we measure the number of genes in the output of RFE that become conditionally independent from the target when conditioning on some subset of genes in the output of UAF. We call this criterion "Relative Conditional Blocking" (RCB). This criterion captures aspects of the *causal bias* of each method. The rationale is the following: Under broad conditions (Spirtes, Glymour, Scheines 2000) if variable X is causing (or is caused by) the target *directly*, there cannot be a non-direct cause or effect variable Y that can render X independent of the target once we condition on Y. Relative conditional blocking measures therefore how close in a causal "directness" sense is the output of RFE to each of the three classification targets compared to the output of UAF (so that the genes that block other genes are causally closer to the target than the blocked genes). Table 4 shows the relative conditional blocking of UAF, and RFE for all classification tasks. We see, for example, that for 31.24% of genes selected by UAF for classification of Adenocarcinomas vs Squamous carcinomas, that there is at least one subset of genes selected by RFE that makes the association of the UAF-selected gene to the target concept “Adenocarcinoma vs Squamous carcinoma”, vanish.

In order to test whether the observed relative blocking ratio (of 31.24% over 100% in the example) can be due to random chance, we perform a permutation test (Noreen 1989) as follows: we simulate 1,000 random pairs of genes with the first set comprising of *j* randomly selected genes and the second set of *i* randomly selected genes from the union of genes outputted by the two methods for a specific task such that *j* is the number selected by RFE and *i* is the number of genes selected by UAF for the examined task. This amounts to a null hypothesis that the two gene selection methods select genes from the same (population of) genes but in different numbers. Then we compute the relative conditional blocking ratio for each random set pair. We derive the empirical distribution of the relative conditional blocking ratio for all 1,000 such randomly selected gene set pairs and examine the probability that the expected ratio under the hypothesis of randomly selected sets is equal or larger to the observed one. The resulting p-values are shown in the last row of Table 4. All observed differences in relative conditional blocking ratios are therefore non-random. A standard cautionary notice with such *conditional* (i.e., as opposed to unconditional) statistical tests (Agresti 1990) is that the results are meaningful only in the context of the specific algorithm’s output (i.e., it does not automatically generalize to *any* output of these algorithms). Furthermore, the *interpretation* of the differences must take into account the differences in the number of selected genes. In the example of Adenocarcinoma vs Squamous carcinoma classification used previously, for every variable *X* output by RFE there is a subset of the output of UAF that is causally closer to the target than *X*.

Cancer vs normal					Adenocarcinomas vs squamous carcinomas				Metastatic vs non-metastatic adenocarcinomas			
Proportion of selected genes	RFE		UAF		RFE		UAF		RFE		UAF	
cardinalities of selected feature sets per data split	12, 6, 6, 3, 6		100, 1000, 25, 100, 25		24, 49, 6, 12, 6		25, 25, 100, 1000, 500		12, 3, 6, 6, 6, 24, 3		500, 1000, 500, 500, 500, 500, 100	
Frequency=1	23	69.70%	1000	79.87%	71	70.30%	1095	64.07%	37	61.67%	1656	38.76%
Frequency=2	7	21.21%	119	9.50%	19	18.81%	472	27.62%	8	13.33%	769	18.00%
Frequency=3	2	6.06%	86	6.87%	8	7.92%	96	5.62%	5	8.33%	610	14.28%
Frequency=4	1	3.03%	33	2.64%	3	2.97%	33	1.93%	4	6.67%	504	11.80%
Frequency=5	0	0.00%	14	1.12%	0	0.00%	13	0.76%	4	6.67%	411	9.62%
Frequency=6	-		-		-		-		1	1.67%	255	5.97%
Frequency=7	-		-		-		-		1	1.67%	67	1.57%
p-value	0.255				0.243				**0.031**			

Table 5. Relative and Absolute Stability of Gene Selection Methods

Another important consideration in feature selection is *stability*, that is, how robust the selected genes are within some method from one data split to another (in cross-validation) and from task to task. Genes that are highly stable across tasks and data splits are less likely to be selected due to sampling variance. Clearly, methods that produce more stable gene sets for a fixed sample size have an advantage over unstable methods and are more likely to point to valuable gene candidates for subsequent experimental research. Table 5 shows the stability of RFE and UAF measured by the following metric: for each one of the two gene selection methods we take the union of all genes selected in some cross-validation split. Then we examine the proportion of genes selected at least 1,2,…,n times (where n is the total number of splits for a task, i.e., 5 or 7). We then compare the two distributions using a G^2 statistic test for independence (Agresti 1990). For example, we see that UAF is more stable than RFE for the metastatic-nonmetastatic classification model since the distribution of multiple occurrences of genes selected by UAF is shifted to higher frequency values relative to RFE and the differences between the distributions of corresponding proportions are statistically significant at the 5% level. In the other two tasks the two methods appear to be equally stable. The lists of genes selected by each method are available to a web supplement to this paper at:
http://discover1.mc.vanderbilt.edu/discover/public/
lungcancerFlairs2003/

Discussion

Our experimental results support the hypothesis that gene expression data combined with powerful learning algorithms can lead to excellent diagnostic models of lung cancer types even with very modest sample sizes and with very low sample-to-feature ratios. These models can distinguish almost perfectly between cancer and normal cells, between squamous carcinomas and adenocarcinomas and between metastatic and non-metastatic adenocarcinomas, all clinically and biologically important states. We found that NNs were not practical for use with all gene predictors but had excellent performance with selected gene sets. KNN, on the other hand, exhibited consistently poorer performance for all classification tasks relative to SVM and NN classifiers irrespective of gene selection. Since gene selection methods can significantly reduce the number of necessary predictors, this leads to the expectation that in the future, the delay and cost for obtaining molecular-based diagnostic test results will be much lower than current genome-wide arraying for the studied (and similar) tasks. An important finding is that the selected genes are different among the gene selection methods despite the fact that both gene selection methods produce high-quality diagnostic models with significant reduction in predictor numbers (with RFE selecting genes sets that are very parsimonious compared to UAF).

In the array used in our analyses 10% of oligonucleotides have the same GenBank gene accession number (i.e., corresponding to the same gene or variants, such as splice variants, mutations and polymorphisms). In the reported analyses we treated unique oligonucleotides as unique genes. In additional experiments (not reported here) in which oligonucleotides with same GenBank accession numbers were replaced by their median we verified that: (a) classifier performance was the same as reported here; (b) RFE is more parsimonious than UAF; and (c) UAF blocks higher percentage of RFE-selected genes. Hence our conclusions are robust to these two possible treatments of oligonucleotides as they relate to unique genes.

In general, since selected genetic markers contain the necessary “expression signatures” of important biological states (i.e., cancer, metastasis, etc.) they may provide guidance in experimental investigation of the pathogenesis of lung cancer. Researchers need to interpret results in the context of the inductive biases of each gene selection method before using these results to design expensive and labor-intensive experiments, however. To facilitate this endeavor we introduced a novel method (Relative Conditional Blocking - RCB) for characterizing the relative causal bias of two feature selection methods. Applied to our data RCB suggests that UAF provides a set of genes that appear to be causally closer to the predicted variables than support vector methods. The validity of this hypothesis needs be verified with experimental work with cell lines or model organisms; however it does provide a valuable starting point for experimental exploration.

There are many possible extensions to the basic RCB method presented here. For example, to control for differences in sizes of algorithm outputs a possible approach is to compare the two algorithms by comparing their RCBs relative to random subsets of genes of size equal to the output size of each algorithm separately. We are in the process of extending the RCB method and systematically studying its properties using synthetic data under different conditions and null hypothesis types. We are also exploring algorithms in which RCB is encapsulated as a stand-alone gene selection and causal hypothesis generation method, returning a Markov Boundary or direct causal neighbourhood of the target variable (for preliminary results, see Tsamardinos Aliferis, and Statnikov 2003b).

References

Agresti, A., Categorical data analysis. John Wiley and Sons; 1990.

Bhattacharjee, A., et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001. 98(24): 13790-5.

Chang C.C., Lin, C.J, LIBSVM: a library for support vector machines (version 2.3). National Taiwan University.

DeLong E., et al. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics 44, 837-845, 1998 .

Demuth, H. and M. Beale, Neural network toolbox user's guide. Matlab user's guide. 2001: The MathWorks Inc.

Duda, R.O., P.E. Hart, and D.G. Stork, Pattern Classification. Second ed. 2001: John Wiley and Sons.

Furey T.S., et al. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics. 2000, 16(10):906-914.

Guyon, I., et al., Gene selection for cancer classification using support vector machines. Machine Learning, 2002, 46: 389-422.

Hagan, M.T., H.B. Demuth, and M.H. Beale, Neural network design. PWS Publishing; 1996.

Noreen E.W. Computer Intensive Methods For Testing Hypotheses. John Wiley and Sons, 1989.

Pagano M. et al. Principles of Biostatistics, Duxbury Thompson Learning, 2000.

Parker, S.L., et al., Cancer statistics, 1996. Cancer J Clin, 1996. 46(1): 5-27.

Provost F., T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing inductionalgorithms. In Proc. Fifteenth Intl. Conf. Machine Learning,

Scholkopf, B., C.J.C. Burges, and A.J. Smola, eds. Advances in kernel methods: support vector learning. The MIT Press; 1999.

Spirtes, P., C. Glymour, and R. Scheines, Causation, Prediction, and Search. Second ed. 2000, Cambridge, Massachusetts, London, England: The MIT Press.

Tsamardinos I, C.F. Aliferis. Towards Principled Feature Selection: Relevancy, Filters, and Wrappers. Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, January, 2003

Tsamardinos I, C.F. Aliferis, A. Statnikov. Algorithms for Large Scale Markov Blanket Discovery. The 16th International FLAIRS Conference, St. Augustine, Florida, USA, May 2003.